Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082593

ARTICLE

Optimizing Forecast Accuracy in Photovoltaic System with Hybrid Artificial Intelligence Model

Yasemin Onal*

Electrical and Electronic Engineering Department, Engineering Faculty, Bilecik Seyh Edebali University, Bilecik, Turkey

*Corresponding Author: Yasemin Onal. Email: yasemin.onal@bilecik.edu.tr

Received: 18 March 2026; Accepted: 02 May 2026; Published: 23 July 2026

ABSTRACT: Photovoltaic (PV) power generation exhibits considerable sensitivity to both weather variability and fluctuations in solar irradiance. Consequently, precise forecasting of PV power is crucial for ensuring grid reliability, load balancing, and the effective functioning of energy markets within a grid-connected solar plant. Conventional forecasting methodologies frequently prove inadequate in accurately capturing the nonlinear and intricate temporal patterns present within PV datasets. To address these shortcomings, this research presents a hybrid short-term PV power forecasting model. This model integrates Neighborhood Component Analysis (NCA) for dimensionality reduction with a Convolutional Neural Network (CNN)-Long Short-Term Memory (LSTM) framework. NCA helps reduce computational complexity while still preserving the important features of high-dimensional PV data. CNN layers are designed to extract spatial features that are localized. In contrast, the LSTM component is adept at capturing temporal dependencies. This combination allows the model to effectively process short-term dynamics. The hybrid model under consideration was evaluated using empirical data obtained from a solar power plant, and its performance was compared to that of conventional machine learning models and individual deep learning (DL) models. The evaluation of performance revealed a MAE of 0.1945, a MAPE of 7.1118, a RMSE of 0.3645, and an R^2 value of 0.976, thereby confirming its enhanced predictive abilities. These results indicate that the hybrid model improves accuracy, precision, and stability in the forecasting of PV power. Consequently, this enhancement underscores the potential of a hybrid DL model to optimize real-time energy management within renewable power systems.

KEYWORDS: Photovoltaic power forecasting; hybrid artificial intelligence; energy decision-making; deep learning; renewable energy systems

1 Introduction

In recent years, renewable energy has become one of the most important alternatives to the environmental and economic problems caused by dependence on fossil fuels. Among these resources, solar photovoltaic (PV) energy stands out due to its rapid increase in installations and wide range of applications. By 2030, global PV installed capacity is expected to reach 1582.9 GW, a significant increase from the 593.9 GW installed in 2019. This growth is largely due to investments in China, India, Germany, the United States, and Japan [1]. Integrated solutions that combine agricultural production with photovoltaic (PV) energy generation in the same area are becoming a sustainable approach for both energy and food security. However, the widespread use of grid-connected PV systems presents operational challenges due to the changing nature of solar radiation and weather. Variations in PV power output can affect both grid stability and agricultural practices. Therefore, continuous monitoring and control are necessary for reliable and efficient operation [2].

Accurate and reliable forecasting of short-term PV power generation is essential for power management and alternative current (AC) grid reliability. A reliable and cost effective predicting model is the most crucial component in connecting large scale PV system power to electrical grids. Improving the accuracy of forecast models improves the operational performance and reliability of PV power generation. Accurate forecasting of PV power generation is crucially dependent on the environmental conditions. Fluctuations in PV power output occur due to changes in solar irradiation throughout the day. This intermittent solar irradiation makes it difficult to trust the PV system power [3].

PV power forecasting models are generally classified into four groups as Physical, Statistical, Machine Learning (ML), and Hybrid Learning (HL) models. Physical prediction models are predicated on mathematical equations that characterize the physical properties of solar radiation and PV components, accounting for diverse environmental factors [4]. These models necessitate detailed input data, including location specifics, meteorological data, and PV system parameters, for practical implementation. Outputs from numerical weather forecast models and sky imaging data are often used to predict future PV power. However, they require detailed input data and are computationally quite complex. Their reliance on numerical weather forecasts, which are generally available in the short term, limits their applicability to short-term forecasts [5,6]. The effectiveness of physical models is further reduced by the availability of low-resolution environmental data and the rapid change in weather events. Statistical prediction models, which utilize historical data and probabilistic methodologies, are employed to model time series by delineating correlations between past and forthcoming values [7]. Their widespread application stems from their straightforwardness and versatility across diverse geographical areas. Furthermore, the effectiveness of statistical models is highly sensitive to changes in the time scale and the granularity of the input data [8]. The Seasonal Autoregressive Integrated Moving Average (SARIMA) [9] and Autoregressive Integrated Moving Average (ARIMA) [10] models are the most popular and often used statistical learning models. ML models are progressively integrated into PV power forecasting, encompassing phases such as data preparation, model training, and prediction. Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) are widely used methods [11,12]. However, these models often struggle to fully represent the complex nonlinear relationships and time-based dependencies found in PV data.

Recent studies increasingly demonstrate improved effectiveness in time series forecasting by modeling complex temporal patterns and long-range dependencies. The focus is on advanced modeling techniques, particularly transformer-based and attention-focused architectures. Unlike traditional iterative models, these methods facilitate parallel processing and the capture of temporal dependencies by utilizing self-attention mechanisms [13–15]. Despite their strong predictive performance, transformer-based models require large scale datasets and significant computational resources for effective training. This constraint could potentially diminish the practical utility of the approach within actual PV systems, given that the datasets typically available are often characterized by restricted sizes and may be subject to noise, incomplete data, and environmental fluctuations. Conversely, hybrid deep learning (DL) models, especially those integrating convolutional neural networks (CNNs) and long short-term memory (LSTM) architectures, have exhibited strong efficacy in the analysis of nonlinear and nonstationary PV data, even when faced with these limitations [16,17].

HL prediction models, such as LSTM networks [18] and CNN [19], which were initially developed for array modeling and image processing, have been successfully adapted to PV power prediction. The CNN model is particularly effective in extracting local spatial patterns from structured data. However, the LSTM model is suitable for learning temporal dependencies in time series data. These models include a combination of Wave Packet Decomposition (WPD) and LSTM, LSTM and Recurrent Neural Networks (RNN) [20], SVM and RNN which increase prediction accuracy by taking advantage of the strengths of different models [21]. Li et al. proposed a HL model combining wavelet packet decomposition and LSTM in the study. In the DL model, WPD is used to decompose the original PV power series into subseries for one hour ahead PV power prediction with five-minute intervals. Four independent LSTM networks are developed for these subseries, the predicted results by each LSTM network are reconstructed, and a linear weighting method is used to obtain the final prediction results [18]. Tang et al. proposed an HL model combining an attention-based extended CNN with bidirectional LSTM (BiLSTM) and further enhanced with a transfer learning strategy. The model is divided into three main stages: input reconstruction, feature extraction, and feature matching. This modular design allows for targeted optimization of individual components rather than adjustments across the entire architecture. The performance of the proposed model was validated using real data from a PV system, demonstrating the effectiveness of the integrated transfer learning strategy [22]. In this regard, Chahboun and Maaroufi used Principal Component Analysis (PCA) to evaluate the performance of various ML models in PV power forecasting [23]. In their analysis, they highlighted the strengths and weaknesses of each method using Bayesian regulated neural networks, elastic network models, SVM, and random forest algorithms. Similarly, Kazem et al. analyzed a 1.4 kW grid-connected PV system using two neural network models full recurrent neural network and PCA-based models, utilizing one year of experimental data. PCA transforms original features into a reduced set of linear combinations. However, it has significant limitations in terms of interpretability. In particular, the transformed features lose the physical meaning associated with the original variables. This makes it difficult to interpret the results meaningfully. Furthermore, PCA does not use class label information even if it is available. This limits its applicability in supervised learning scenarios [24]. Due to the large dimensionality and complex structure of PV data, it is necessary to use more advanced dimensionality reduction techniques that preserve distinctive information while increasing computational efficiency in the prediction model. In this respect, Neighborhood Component Analysis (NCA) is a more suitable alternative because it performs feature selection without significant information loss.

Recent studies published in 2025 further confirm the rapid progress of photovoltaic power forecasting. Recent work has explored transferable deep learning frameworks for multi-country PV forecasting, probabilistic solar power forecasting models, updated reviews of machine-learning-based PV prediction, and transformer-based transferable architectures for data-limited PV sites [15,25]. These studies indicate that the field is moving toward more adaptive and scalable forecasting frameworks, while also showing that model complexity and data requirements remain important practical considerations [13]. This study presents the NCA–CNN–LSTM framework, which aims to capitalize on these benefits by integrating supervised feature selection via NCA. This approach diminishes input dimensionality, thereby improving model efficiency. Unlike current methods that focus on attention mechanisms, the proposed model offers a beneficial balance between prediction accuracy, computational efficiency, and stability, particularly when applied to moderately sized PV datasets.

The main scientific contribution of this study is the integration of NCA-based supervised feature transformation with CNN-LSTM architecture for short-term PV power forecasting. This allows for improved feature relevance, reduced redundancy, and consistently good performance under limited data conditions. The contributions of this study can be summarized as follows: (i) A novel supervised feature algorithm based on NCA to eliminate redundant inputs before DL has been integrated into a CNN-LSTM architecture. (ii) Unlike traditional hybrid models using raw or unsupervised representations, the proposed model has improved convergence stability and generalization capability through feature weighting. (iii) The model has been validated by extensive benchmarking against multiple ML and DL baseline models under the same data pods. (iv) The prediction mechanism's explainability was established via SHAP, ICE, and PDP analyses, thereby offering physical interpretability. The model's suitability for real-time PV system operation has been confirmed.

This paper introduces a new hybrid DL model. It combines NCA, CNN, and LSTM networks for accurate power estimation in PV systems. To the best of the author's knowledge, there are a limited number of studies that have examined the integration of NCA-based supervised feature transformation with CNN-LSTM architecture for short-term PV power forecasting. The experimental analysis was conducted using PV system data, including irradiance, ambient, and module temperature parameters. NCA is used in the model to automatically identify and extract the most important features from the input data. These extracted features are then used by the CNN-LSTM estimator, which has the power of prediction and pattern recognition, to provide precise and systematic power estimation. The proposed model's performance is evaluated using four different error metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). The proposed hybrid DL model is compared with traditional ML and other DL models. The proposed NCA-CNN-LSTM framework demonstrates remarkable potential and adaptability, providing a solid foundation for PV system power estimation. This paper's remaining sections are organized as follows: The materials and procedures are described in full in Section 2, together with the dataset description, data preparation, performance evaluation measures, the proposed NCA-CNN-LSTM forecasting methods, including NCA, CNN algorithm and LSTM algorithm. Comparative evaluation of the performance of the proposed model with ML models and DL models is given in Section 3. SHAP, ICE and PDP analyses are given in Section 4. Finally, conclusions are provided in Section 5.

2 Materials and Methods

2.1 Description and Preparation of the Dataset

Solar energy forecasting essentially involves predicting solar energy production for upcoming time intervals based on predicted weather parameters, including irradiance, ambient temperature, humidity, wind speed, and other relevant factors. The dataset used in this study was obtained from a publicly available Kaggle dataset [26]. Therefore, no physical experimental materials were used. Data were collected over a 34-day period from a grid-connected PV power plant near Nasik, Maharashtra, India. Power generation datasets were collected at the inverter level, with multiple solar panel lines connected to each inverter. Sensor data were collected at the plant level, via a single array of sensors optimally placed within the plant. The data consists of a power generation dataset and a sensor dataset. The dataset was divided into three subsets: 80% for the training set, 20% for the validation set, and the last three days were allocated as a test set to preserve the temporal structure of the data and provide realistic forecasting conditions. This forward validation strategy is particularly suitable for time series forecasting problems [27].

The dataset used in this study includes measurements from 22 inverters, collected every 15 min from May 15 to June 17, 2020. To create forecasts, the data were resampled to an hourly frequency of 60 min. At the same time, the original 15-min data were kept for exploratory data analysis. In the first part of this study, the exploratory data analysis is performed on the PV energy system dataset. Data preparation is performed to clean and transform the data, and feature engineering is applied to select or generate relevant features to train the model. All inverter measurements were aligned on a unified temporal axis. Missing data were handled based on production periods and completed using interpolation techniques for both production and non-production intervals. Following preprocessing, the dataset was structured as time-series sequences, where each sample represents consecutive temporal observations. This architectural design facilitates the model's capacity to accurately represent the temporal relationships inherent in the input variables and the desired output. The dataset encompasses crucial meteorological and electrical parameters, including daily solar irradiation (W/m^2), ambient temperature (°C), module temperature (°C), daily DC electricity generation (W), grid AC electricity consumption (kW), daily power yield (kW/h), and total power yield (kW/h). Therefore, this setup offers sufficient time resolution for short-term PV power forecasting, while also maintaining temporal consistency during the model's training. Evaluation criteria are used to evaluate the performance of the proposed model on the test data.

Fig. 1 provides a comprehensive analysis of the PV system through multiple subgraphs. Fig. 1a shows the daily variations of AC power, DC power and irradiation, showing a characteristic bell-shaped curve that shows solar energy production throughout the day. It is seen that AC and DC power production vary in parallel with solar radiation throughout the day. Power production is also low due to low morning temperatures. AC and DC power production peaks when the irradiance in the cross-section reaches its maximum. In the evening, power production drops again due to the structure of the irradiation. Fig. 1b shows the daily yield, reflecting energy output trends over time. The daily energy production of a PV system varies over time. Fluctuations can arise from parameters such as weather conditions, cloudiness, and system changes. It is generally observed that production decreases during certain periods, while it is more concentrated during other periods. Fig. 1c compares the ambient and module temperatures, showing how the module temperature rises significantly above ambient levels during the peak sunlight hours. The solar irradiation increases at the module temperatures, and the feeding rates reach their maximum levels. High module temperatures can reduce power output because of the negative changes in PV characteristics and the resulting temperature effects. Finally, Fig. 1d provides a bar graph representation of the total daily efficiency for each day. The total daily energy production varies. On some days, production is quite high, while on other days it is intermittent. Weather, cloud cover, and other specific factors, along with potential systems, can all affect this. These visualizations, when considered together, provide information about how well the system works, its environmental impact, and the general patterns of power generation.

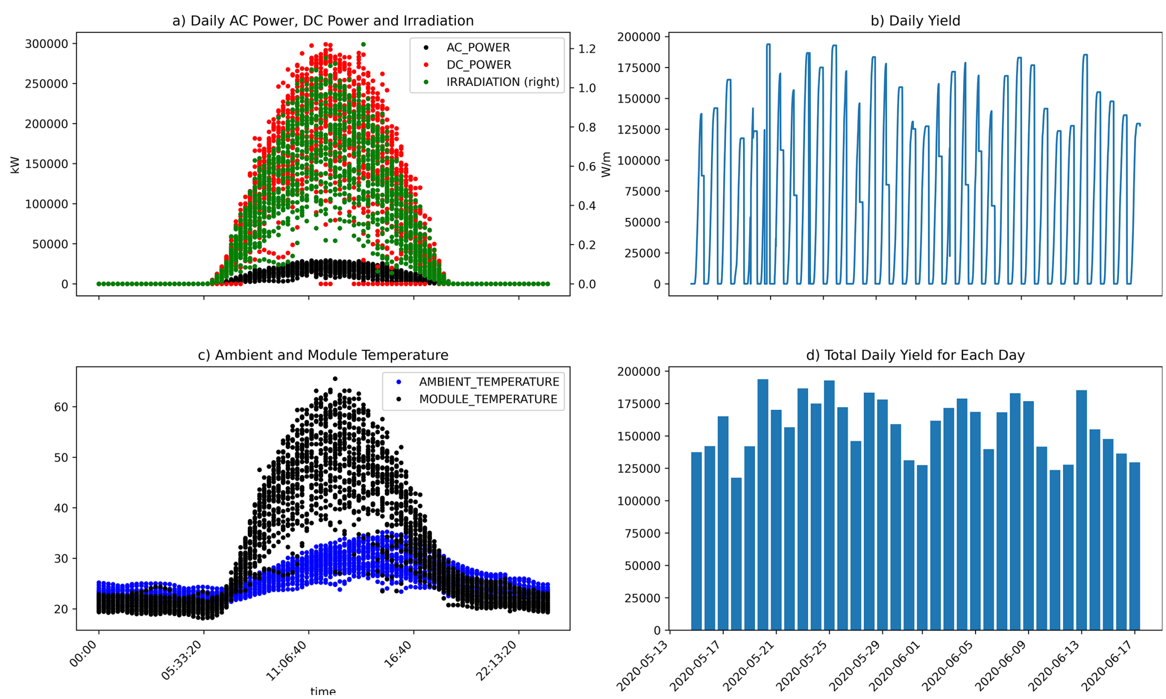


Figure 1: Analysis of PV system performance: (**a**) daily variation of AC/DC power generation and irradiance, (**b**) daily yield production, (**c**) comparison of ambient and module temperatures, (**d**) variation of total daily yield production over time.

2.2 Performance Evaluation Metrics

The process of estimating energy production follows a structured method. First, NCA is used to automatically find the most important input features. Then, the selected features are preprocessed through normalization. This ensures consistent scaling and improves the efficiency of the learning process. Model performance is evaluated using widely adopted statistical error metrics, including MAE, MAPE, RMSE, and R^2 given in Eqs. (1)–(4) [28].

These metrics offer distinct viewpoints on prediction accuracy, robustness, and sensitivity to significant errors. MAE and MAPE assess average absolute and relative deviations, respectively. RMSE, conversely, places greater emphasis on larger errors, and R^2 quantifies the proportion of variance in the target variable that the model accounts for. Their computational efficiency and widespread application in existing research render them appropriate for both large-scale and real-time applications, facilitating dependable benchmarking against prior investigations. Consequently, lower MAE, MAPE, and RMSE values, coupled with higher R^2 values, signify enhanced predictive performance [29].

$$MAE = \sum_{J=1}^{n} \frac{\left|y_{j,act} - y_{j,Pre}\right|}{n} \tag{1}$$

$$MAPE = \frac{1}{n} \sum_{J=1}^{n} \frac{\left|y_{j,act} - y_{j,Pre}\right|}{y_{j,act}} * 100 \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{J=1}^{n} \left(y_{j,act} - y_{j,Pre}\right)^2}{n}} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \left(y_{j,act} - y_{j,Pre}\right)^{\ 2}}{\sum_{J=1}^{n} \left(y_{j,act} - \overline{y_{j,act}}\right)^2} \tag{4}$$

where n is the number of data points over which the predicted errors are assessed, $y_{j,act}$ and $y_{j,Pre}$ stand for the hybrid model actual and estimated output power values.

2.3 Proposed NCA-CNN-LSTM Forecasting Methods

To enhance prediction accuracy, the proposed hybrid NCA-CNN-LSTM framework integrates supervised feature selection with DL–based spatial and temporal modeling. The overall architecture of the method is illustrated in Fig. 2. Initially, PV power generation data are combined with sensor measurements and preprocessed through data cleaning, resampling, and normalization procedures to ensure consistency and numerical stability. To ensure data integrity, the initial step involved the removal of missing and inconsistent values. Following this, the data was resampled to a uniform temporal resolution, thus promoting consistency. Finally, min-max normalization was applied to scale all input features, which in turn enhanced both the model's convergence speed and its overall stability. After preprocessing, NCA is then applied to identify the most discriminative input features and reduce dimensionality while preserving critical relationships among variables. This step reduces redundancy and enhances the relevance of the input variables prior to deep learning.

After preprocessing, the dataset is divided into training, validation, and test subsets using a forward validation approach. The modified dataset is then used as input for a one-dimensional CNN–LSTM architecture. Conv1D and MaxPooling1D layers are utilized to extract local temporal features and diminish dimensionality, thereby retaining the most significant patterns [18]. The resulting feature maps are subsequently processed by an LSTM layer, which captures sequential dependencies within the learned feature space. In this architecture, the LSTM layer operates on the feature maps generated by the convolutional layer, enabling the model to learn relationships within the structured input representation. After that, the resulting feature maps are flattened and sent to a dense layer, which performs nonlinear regression. Ultimately, Flatten and Dense layers transform the acquired representations into higher-level features, and the output layer produces continuous PV power predictions.

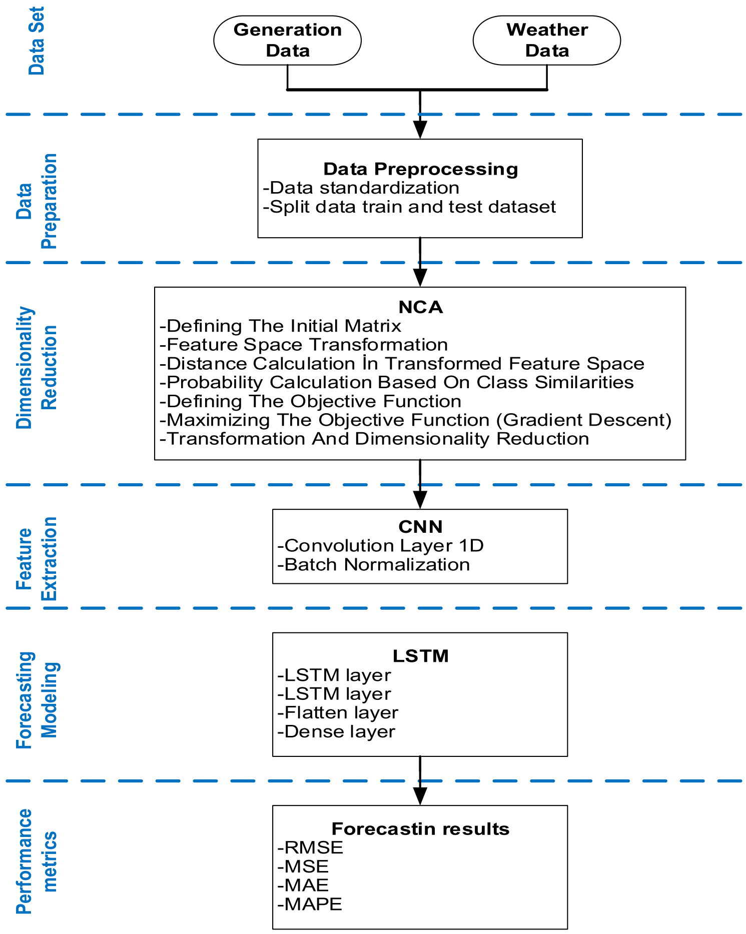


Figure 2: The proposed NCA-CNN-LSTM architecture.

The proposed NCA–CNN–LSTM model effectively captures relationships in multivariate sensor data, including solar irradiation, ambient temperature, and module temperature, while also modeling temporal dependencies critical for AC power prediction. Feature extraction is performed using a Conv1D layer with 128 filters and 2 kernel dimensions. Then, Maxpooling is performed. Temporal dynamics were modeled using LSTM layers with 128 and 64 hidden units, allowing the model to learn patterns over both short and long periods. The output layer consists of a single neuron with linear activation, producing the predicted PV power value. The model is trained using the Adam optimization algorithm with a learning rate of 0.001, a batch size of 32, and 150 epochs. These hyperparameters are selected experimentally to balance model accuracy and computational efficiency. Overall, the integration of NCA with the CNN-LSTM architecture is key. Table 1 gives the configuration parameters of the proposed model. Alternative configurations were also tested, and the selected model achieved the best RMSE and MAE performance.

Table 1: Configuration parameters of proposed model.

Layer	Parameters	Activation	Additional Notes
Conv1D	Filters: 128, Kernel size: 2	ReLu	Input shape: (3, 1)
MaxPooling1D	Pool size: 2	–	–
Flatten	–	–	–
LSTM	Hidden units: 128	ReLu	Return_sequences: True
LSTM	Hidden units: 64	ReLu	Return_sequences: False
Output Dense	Units: 1	Linear	Final prediction layer

2.4 Neighborhood Component Analysis NCA for Dimensionality Reduction

NCA, a supervised dimensionality reduction method, was initially conceived to enhance the efficacy of k-nearest neighbor classification by refining class within a transformed feature space. NCA is adapted for regression-based PV power forecasting as a feature transformation and weighting technique. Since NCA requires discrete target labels for supervised transformation, the continuous PV power output is discretized into quantile-based groups to construct pseudo-class labels. This discretization preserves neighborhood relationships among samples with similar output values, thereby enabling effective supervised feature learning. Similar discretization-based adaptations have been successfully employed in previous studies to extend classification-oriented feature learning techniques to regression problems [30]. Importantly, this transformation is applied only during the feature extraction stage, while the final CNN–LSTM model is trained using the original continuous PV power values. Therefore, the regression objective of the forecasting model remains unaffected. Accordingly, the decomposed target groups are treated as class labels in the feature transformation stage. The standard classification-based formulationof NCA presented in Eqs. (5)–(10) is used directly in this study. This formulation ensures that the classification-based NCA framework remains theoretically consistent when applied to discretized regression targets. NCA operates by learning a linear transformation matrix A that maps each input vector x_i to a new feature representation:

$$y_i = Ax_i \tag{5}$$

The probability p_{ij} that a data point i is the neighbor of j is computed as follows:

$$p_{ij} = \frac{\exp(-\left\|y_i - y_j\right\|^2)}{\sum_{k\neq i} exp(-\left\|y_i - y_k\right\|^2} \tag{6}$$

here $\left\|y_i - y_j\right\|$ represents the Euclidean distance between transformed feature vectors y_i and y_j. The likelihood that a given point x_i is correctly classified is determined by:

$$p_i = \sum_{j\in C_i} p_{ij} \tag{7}$$

where C_i denotes the set of data points belonging to the same class as x_i. The algorithm seeks to maximize the expected number of correctly classified samples, represented by the objective function:

$$\mathrm{L}\left(A\right) = \sum_{i} p_i \tag{8}$$

The gradient of the objective function with respect to the transformation matrix A is computed as:

$$\frac{\partial \mathrm{L}\left(A\right)}{\partial A} = \sum_{i}\sum_{j}\left(p_{ij} - \delta_{ij}\right)\left(y_i - y_j\right)x_i \tag{9}$$

where δ_{ij} is an indicator function that equals 1 if x_i belong to the same class and 0 otherwise. Using gradient ascent, the transformation matrix is iteratively updated:

$$A \leftarrow A + \eta \frac{\partial \mathrm{L}(A)}{\partial A} \tag{10}$$

where η signifies the learning rate, which governs the magnitude of the update's step. The optimization procedure iterates until the objective function $\mathrm{L}(A)$ achieves convergence, a state defined by the insignificance of successive improvements. The incorporation of NCA within CNN–LSTM architectures offers multiple benefits compared to traditional machine learning methodologies. Through the assignment of greater significance to pertinent features and the suppression of superfluous or noisy variables, NCA effectively diminishes the input space's dimensionality, thereby lessening the likelihood of overfitting. This enhanced feature representation facilitates the subsequent CNN and LSTM layers in more effectively discerning intricate patterns within the data. In particular, the Conv1D layer captures localized feature interactions, while the LSTM layer models sequential dependencies within the transformed feature space [31]. The proposed NCA–CNN–LSTM model improves both its ability to generalize and its computational efficiency by focusing on informative feature interactions, unlike CNN models that use the entire set of features.

2.5 Convolutional Neural Networks CNN Algorithm

CNNs were initially designed for image classification tasks. These networks use convolutional filters to extract hierarchical and discriminative feature patterns. This is done through trainable weights, biases, and nonlinear activation functions. As a supervised learning model, CNNs follow a structured process that includes feature extraction, training, evaluation, and prediction. Compared with other deep learning techniques, CNNs automatically learn hierarchical feature representations. This generally requires limited preprocessing, resulting in an efficient modeling framework [32]. The steps of the CNN algorithm are summarized in the pseudocode shown in Fig. 3.

```
# Step 1:CNN Architecture:
        Initialize CNN model, Add Convolutional Layer(s) with filters, kernel size,
        activation function, Add Pooling Layer(s) to reduce dimensionality,Flatten the
        feature maps into a vector, Add Fully Connected Layer(s), Add Output Layer
        with activation function (Softmax for classification).
# Step 2: Compile the Model:
        Define loss function (categorical cross-entropy), Define optimizer (Adam or
        SGD), Specify evaluation metrics (accuracy).
# Step 3: Train the Model: For each epoch:
        Forward propagate input data through CNN, Compute loss using the loss
        function, Backpropagate errors and update weights using the optimizer,
        Validate the model on validation dataset.
# Step 4: Evaluate the Model:
        Use test dataset to assess model performance, Calculate accuracy, precision,
        recall, and loss.
```

Figure 3: The pseudocode of CNN algorithm.

Due to the large number of parameters that can be adjusted, CNNs are prone to overfitting, especially when the training data is small or has a lot of redundancy [33]. To improve their ability to generalize, regularization methods like dropout and early stopping are often used. Dropout randomly deactivates neurons during training, encouraging the network to learn robust feature representations, while early stopping prevents excessive training [34]. A typical CNN architecture consists of an input layer followed by convolutional and pooling layers for hierarchical feature extraction and dimensionality reduction, respectively. After that, the extracted features are flattened and then sent to fully connected and output layers, which leads to the final prediction.

2.6 Long Short-Term Memory LSTM Algorithm

LSTM networks, a sophisticated variant of RNN, are purpose-built for processing sequential data and effectively modeling long-range dependencies. Conventional RNNs frequently encounter difficulties stemming from the vanishing gradient problem, thereby restricting their capacity to retain information across extended sequences. LSTM circumvents this limitation through the implementation of a gated architecture. This architecture uses forget, input, and output gates, which control how information flows through the network. As a result, LSTMs can remove unnecessary or irrelevant data while keeping important information. At each time step, an LSTM processes the current input along with the previous hidden state and cell state, representing short-term and long-term memory [35]. The forget gate determines which information from the previous cell state should be retained. The input gate controls the inclusion of new information into the memory cell. Candidate information is generated using a tanh activation function. It is selectively updated via a sigmoid-regulated mechanism [36]. The output gate then combines the updated cell state with the current input to produce the hidden state. It provides both network output and short-term memory for the next time step. LSTM networks are specifically designed to overcome the vanishing gradient problem and effectively capture temporal dependencies in sequential data [37].

3 Results

This segment assesses the efficacy of the suggested hybrid NCA-CNN-LSTM model in comparison to established ML and DL models, specifically LSTM, CNN, and CNN-LSTM architectures. Model performance was assessed through the application of R^2, MSE, and MAE metrics. The experiments were carried out using Python, employing established ML and DL libraries. To ensure the findings can be reproduced, the code used in this study is available upon request.

3.1 Data Analysis

Table 2 presents a statistical summary of the solar data, including mean, standard deviation, and outliers. The AC power generation process initiates roughly at 6 am, demonstrating a progressive escalation until it peaks between 10 am and 2 pm, with average values exceeding 1200 W. Following this, a decrease toward zero is observable around 7 pm, reflecting the natural solar cycle. It is important to note that the most significant values were observed during the early morning and late afternoon hours. These values are attributed to weather variability or shading effects.

Fig. 4 presents the correlation heat map between key PV system variables. The strongest connection, a perfect correlation of 1.0, is seen with AC power, highlighting its crucial role in energy production. AC power also shows strong correlations with module temperature (0.96) and ambient temperature (0.73). These values provide important information regarding data variability for subsequent modeling. Variables such as daily and total efficiency were excluded from the dataset as they showed weaker correlations. Despite exhibiting an excellent correlation with AC power and a correlation coefficient of 1.0, DC power was excluded from the dataset to prevent data leakage and allow for a more realistic assessment of prediction performance.

Table 2: Quantitative form of input data used in this study.

Column	Count	Mean	Std. dev.	Min	Max	25%	50%	75%
DC Power	3183	3094.17	3996.26	0.00	1.35×10^4	0.00	343.15	6337.45
AC Power	3183	302.59	390.49	0.00	1.32×10^3	0.00	33.16	620.83
Daily yield	3183	7.18×10^4	6.59×10^4	0.00	1.93×10^5	96.02	6.60×10^4	1.29×10^5
Total yield	3183	1.50×10^8	1.70×10^7	0.00	1.56×10^8	1.52×10^8	1.53×10^8	1.54×10^8
Ambient temp.	3183	25.53	3.35	20.39	35.25	22.70	24.61	27.92
Module temp.	3183	31.09	12.26	18.14	65.54	21.09	24.62	41.33
Irrad.	3183	0.22	0.30	0.00	1.22	0.00	0.02	0.44

Figure 4: Correlation matrix for the features in the dataset.

Fig. 5 presents hex contour plots that show the relationship between AC power output and important environmental factors: solar irradiation, module temperature, and ambient temperature. Fig. 5a shows a clear positive correlation between AC power and solar irradiation. The concentrated clustering along the diagonal slope reinforces the direct relationship between these two variables. In contrast, Fig. 5b,c illustrates the relationship between AC power and ambient temperature and module temperature. Compared to irradiation, the data points are more scattered, showing a weaker and less consistent relationship compared to solar irradiation. Edge histograms show that most observations occurred in low irradiation and ambient temperature ranges. The strong, direct relationship between solar irradiation and AC power is clearly visible, confirming the direct link between them. Furthermore, the observed positive correlation between PV module temperature and the surrounding air temperature demonstrates complex interactions influenced by environmental factors.

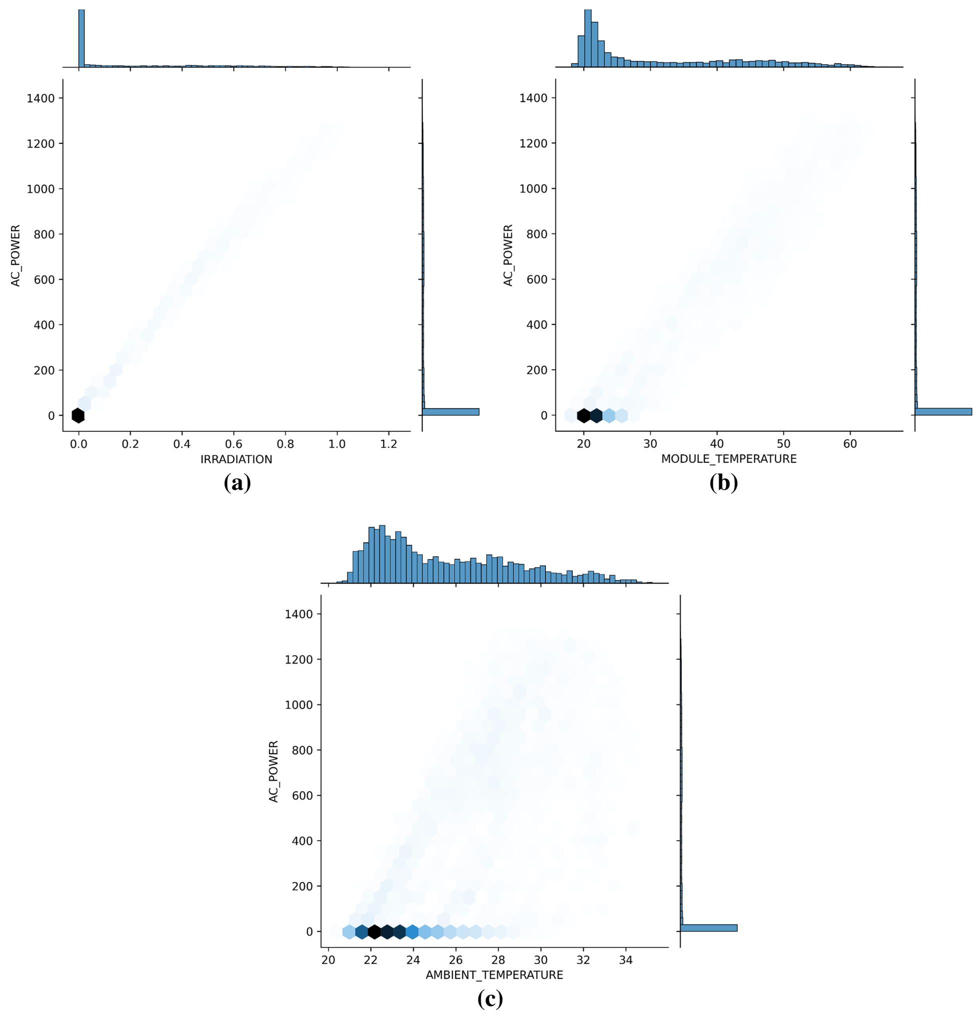


Figure 5: Visualization of input–output parameter relationships using hexagonal contour graphs: (**a**) irradiation, (**b**) module temperature, and (**c**) ambient temperature.

3.2 Comparative Evaluation of Machine Learning and Deep Learning Models

For all models, forward validation is applied to generate predictions for the next data point. This validation strategy is particularly suitable for time series forecasting because it reflects the practical scenario where forecast accuracy gradually decreases with increasing forecast horizons. In this section, additional experiments were conducted to verify that the proposed approach outperforms alternative ML and DL models. Five distinct ML models and four DL models were assessed using MAE, MAPE, RMSE, and R^2 metrics.

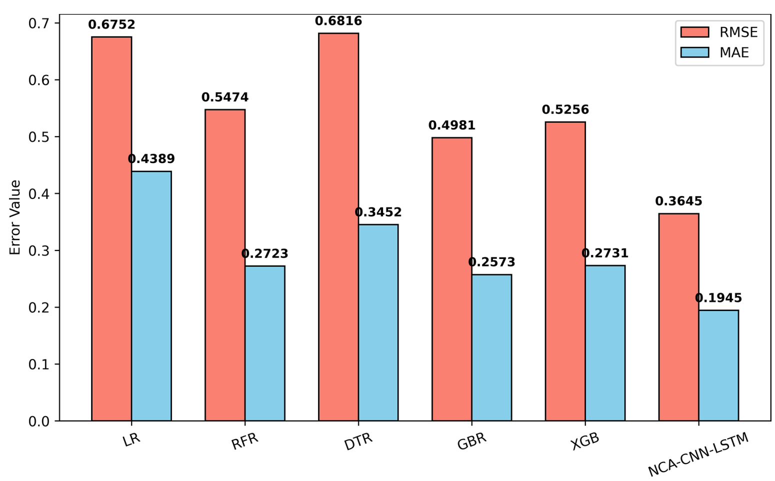
Table 3 presents the numerical results. The NCA–CNN–LSTM model showed the best performance, with the lowest error metrics and the highest R^2 value. Specifically, the model had a MAE of 0.1945 and a RMSE of 0.3645. This model's predictive ability was better than that of the other models. Although the improvement in MAE and RMSE values is small compared to the basic CNN–LSTM model, the results show a consistent improvement in prediction performance. As seen in the analysis, the improvement in RMSE and MAE values confirms that NCA effectively reduces redundant information and improves model robustness.

Table 3: Proposed NCA-CNN-LSTM, ML and DL models performance comparison.

Model	MAE	MAPE	RMSE	R^2
LR	0.4389	52.7296	0.6752	0.819
RFR	0.2723	9.2625	0.5474	0.846
DTR	0.3452	13.5652	0.6816	0.817
GBR	0.2573	11.0111	0.4981	0.856
XGB	0.2731	10.3419	0.5256	0.851
DNN	0.1963	20.75	0.3897	0.860
CNN	0.1958	13.22	0.3750	0.864
LSTM	0.1954	9.0096	0.3715	0.867
CNN-LSTM	0.1949	8.2269	0.3691	0.891
NCA-CNN-LSTM	0.1945	7.1178	0.3645	0.976

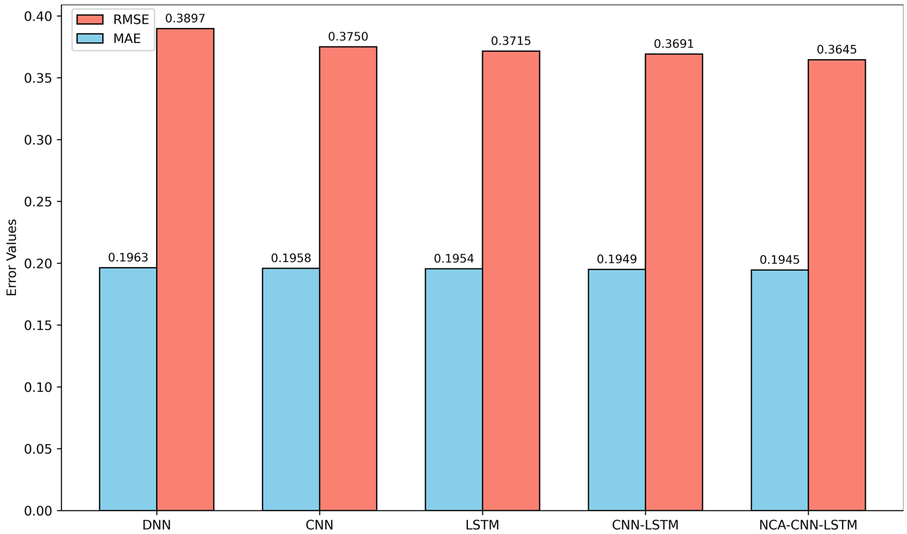
An ablation study was conducted to evaluate the effect of NCA by comparing the proposed model with a basic CNN–LSTM model trained on the original feature set. The results showed that the inclusion of NCA improved feature representation and made model performance more stable. This comparison serves as an ablation analysis that explicitly evaluates the effect of NCA on model performance [38]. It also showed that it reduced input size, computational cost, and convergence time. Fig. 6a,b presents the comparative error analysis of the evaluated ML and DL models in terms of RMSE and MAE.

To assess the statistical significance of the obtained performance differences, the Wilcoxon signed-rank test was applied to the models on absolute prediction error. The results show that the proposed NCA–CNN–LSTM model provides a statistically significant improvement compared to the basic CNN–LSTM model. The test result is given in Table 4 (Wilcoxon test: $p = 1.6185 \times 10^{-4} < 0.05$). It should be noted that the performance difference between the CNN-LSTM and NCA-CNN-LSTM models is 0.0004 MAE. Despite this statistical significance, the absolute improvement in error metrics remains small. However, this improvement was achieved using a reduced and more informative feature set. This demonstrates feature efficiency and increased model stability. Therefore, the contribution of the proposed model, apart from large numerical gains in prediction accuracy, is improved feature definition and model stability.



(a)

Figure 6: (Continued)



(**b**)

Figure 6: Comparative analysis of error values across (**a**) ML models, (**b**) DL models.

Table 4: Statistical significance test result.

Test	Statistic	*p*-Value
Wilcoxon test	15,472.0	1.6185×10^{-4}

4 Discussion

Fig. 7a shows the feature importance analysis graph, and Fig. 7b shows the zoomed-in graph of the input features module temperature and ambient temperature. The graphs illustrate the extent to which each input variable contributes to the prediction of AC power. Radiation was identified as the most influential factor with a score of 0.99496, making it the primary determinant of the model's prediction output. All other features, modulus temperature and ambient temperature, had very low significance scores of 0.00259 and 0.00245, respectively.

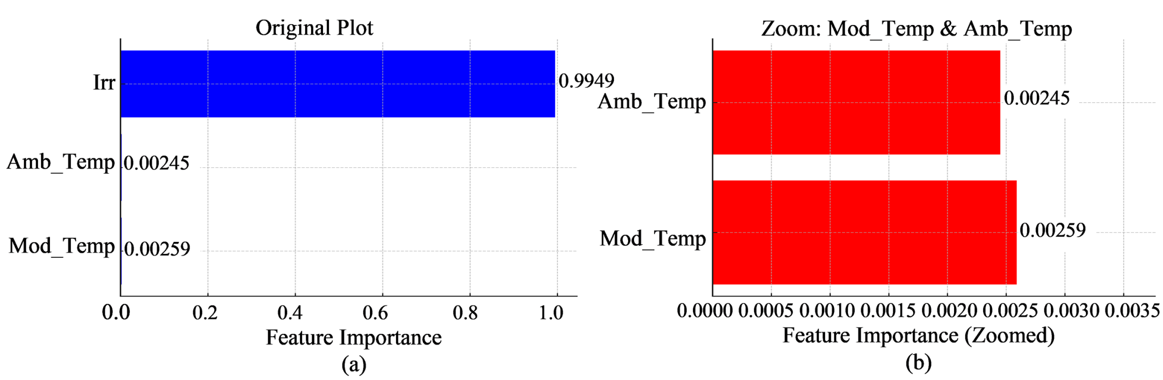


Figure 7: Sklearn permutation importance package for (**a**) feature importance analysis, (**b**) zoom input features.

The SHAP plot shows the relative contribution and impact of input variables on the prediction model's output. As seen in Fig. 8, solar irradiance exhibits the highest influence. Both low and high values have significant effects on the prediction output, indicating its dominant role in the prediction model. Conversely, the influence of module temperature and ambient temperature is comparatively minor, as indicated by SHAP values clustering near zero. The color gradient, transitioning from blue to red, represents the feature values; thus, elevated irradiance values are associated with more pronounced positive effects on the model's output.

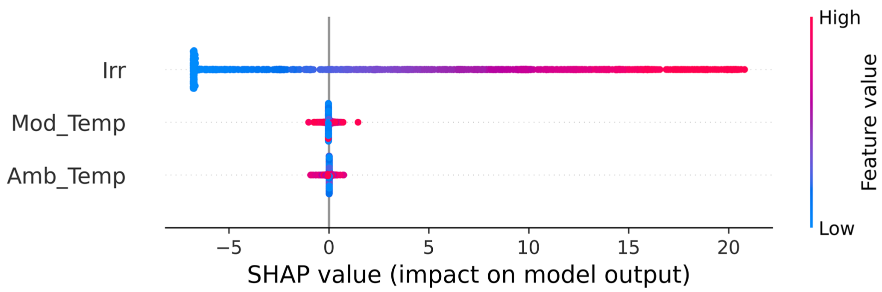


Figure 8: Impact of individual features using the SHAP plot.

After assessing feature importance, ICE and PDP analyses were performed together to investigate the specific effects of the input factors. The results in Fig. 9 show that increasing solar irradiance leads to a proportional increase in AC power prediction (Fig. 9a), while module temperature and ambient temperature contribute marginally negative effects (Fig. 9b,c). It is known that the input variables interact with each other, and these interactions significantly affect the AC output power. Therefore, the effect of these two variables on AC power was examined using a 3D plot.

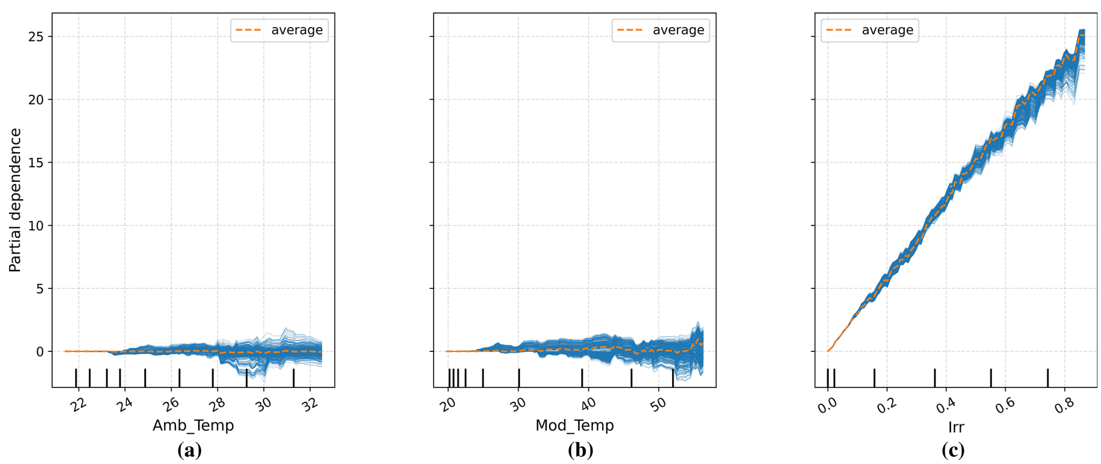


Figure 9: ICE and PDP analysis of the input variables with (**a**) ambient temperature, (**b**) module temperature and (**c**) irradiation values for AC power prediction.

Fig. 10 shows three-dimensional scatter plots that illustrate the relationships between solar irradiation, ambient and module temperatures, and AC power output. Fig. 10a, the relationship between solar irradiation, ambient temperature, and AC power reveals a strong and consistent increase in AC power with increasing irradiation levels. The clustering of data points along a clear trend indicates that irradiation is the dominant factor, while ambient temperature causes only minor changes in power output. Fig. 10b shows the interaction between solar irradiation, module temperature and AC power, and Fig. 10c shows the combined effect of ambient temperature, module temperature and AC power. Although module temperature increases at higher irradiation levels, its direct effect on AC power remains limited. The spatial distribution indicates a less pronounced and indirect correlation between temperature and the observed variables, in contrast to irradiation.

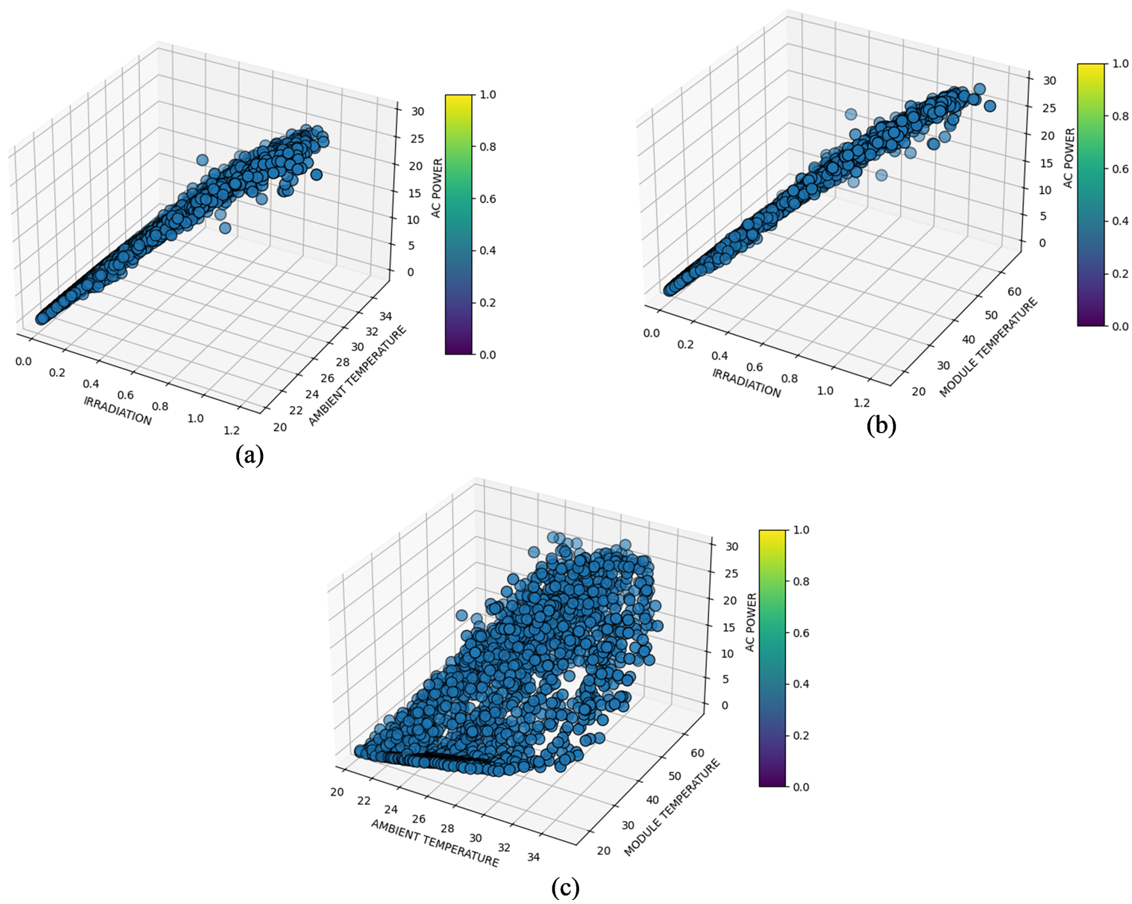


Figure 10: Analysis of partial dependence depicting the relationship between (**a**) solar irradiation, ambient temperature, and AC power, (**b**) solar irradiation, module temperature and AC power, (**c**) ambient temperature, module temperature and AC power.

Furthermore, the color-coded depiction of normalized AC power underscores the variability in output across diverse operational scenarios. This representation corroborates the assertion that irradiation is the principal determinant, while thermal parameters exert a comparatively minor influence. SHAP, ICE, and PDP analyses show that solar irradiation positively influences AC power predictions, while ambient and module temperatures negatively influence them. Fig. 11 shows the RMSE curves obtained in the training, validation, and test datasets, and the RMSE as a function of the forecast horizon for the proposed NCA-CNN-LSTM model. As shown in Fig. 11a, the decrease in RMSE from the training set to the test set suggests that the proposed model converges consistently and generalizes well. The validation and testing phases showed no performance decrease. This indicates that the combination of the proposed NCA-based feature selection and the CNN-LSTM architecture avoids overfitting. According to Fig. 11b, a gradual increase in RMSE is observed for longer forecast horizons.

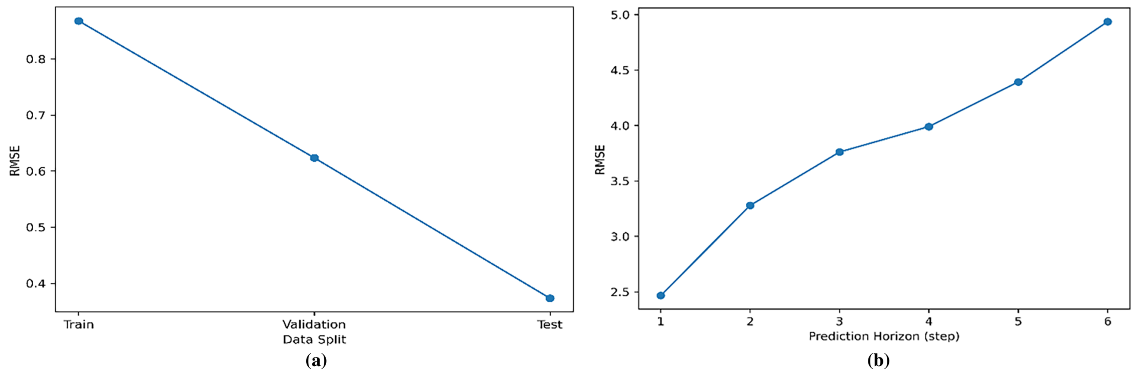


Figure 11: (**a**) Training, validation and test RMSE learning curve, and (**b**) RMSE variation with respect to prediction horizon for the proposed NCA-CNN-LSTM model.

Fig. 12 presents the scatter plot and the histogram of the residuals for predicted and actual AC power values. According to Fig. 12a, the strong linear alignment of data points along the identity line indicates a high level of agreement between the predicted and measured power values across the entire operating range of the PV system. The limited dispersion around the diagonal confirms the consistently better performance and consistency of the model. It also demonstrates its effectiveness in capturing both low- and high-power generation regimes without significant bias. According to Fig. 12b, the residuals are symmetrically distributed around zero. This shows that the model produces unbiased predictions. While most errors are concentrated in a narrow range, only a small number of outliers appear in the tails. This symmetrical and almost Gaussian error distribution demonstrates the robustness of the proposed model under varying operating conditions. It also confirms its ability to accurately capture the nonlinear dynamics of the PV energy generation process. The results confirmed that the improved accuracy came from carefully prioritizing features, not from making the network deeper.

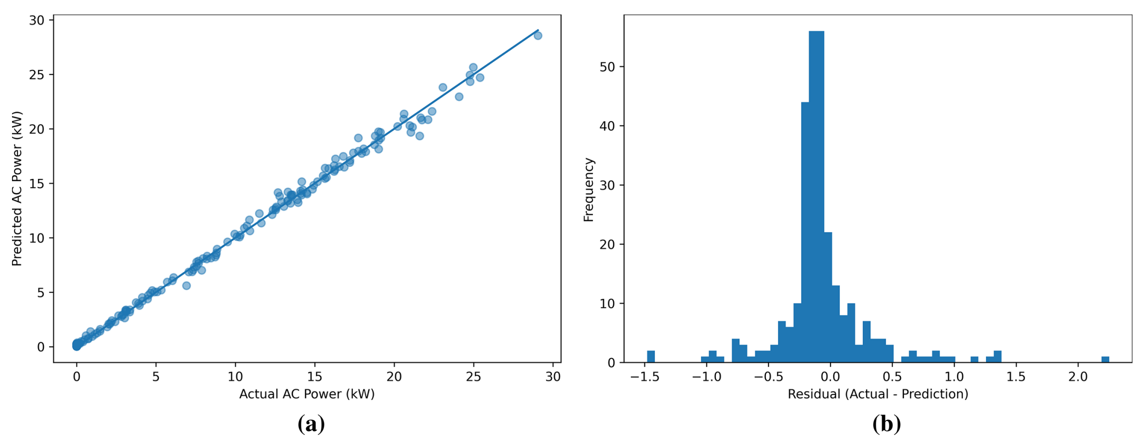


Figure 12: (**a**) Scatter plot, and (**b**) histogram of residuals for predicted and actual AC power values of the proposed NCA-CNN-LSTM model.

The dataset used in this study covers a short period, specifically 34 days, and comes from a single PV system. This choice was made intentionally, reflecting the main goal of short-term PV power forecasting. Here, capturing high-frequency temporal dynamics is more critical than long-term seasonal variability [31]. In such applications, short-term datasets with fine temporal resolution are usually sufficient to model rapid fluctuations in solar irradiation and power output. Furthermore, the dataset consists of real measurements collected from a real grid-connected PV system and obtained from a publicly available source. The reliability of the data and the reproducibility of experimental outcomes are thereby assured, a prerequisite for validating the proposed model within realistic operational parameters [32]. Nevertheless, the model's generalizability across diverse climatic conditions and seasonal fluctuations could be constrained. Consequently, forthcoming research studies will concentrate on augmenting the dataset to encompass multi-location and multi-seasonal data, alongside conducting cross-validation across a range of geographic regions, with the aim of enhancing the model's robustness and practical utility. NCA offers a supervised transformation of the feature space, distinguishing it from traditional feature selection methods such as Recursive Feature Elimination and mutual information. Guyon observed that this methodology allows the model to better identify complex relationships between input variables and the target output. Moreover, the comparative assessment of CNN-LSTM and NCA-CNN-LSTM models serves as an ablation study, thus clarifying the importance of the NCA-induced feature transformation.

5 Conclusions

This research presents a hybrid DL model integrating NCA, CNNs, and LSTM networks for short-term PV power forecasting. The proposed approach addresses the challenges of high dimensionality, nonlinearity, and variability in PV datasets by incorporating supervised feature transformation. A key advancement of this model is its incorporation of supervised feature selection through NCA, a technique that effectively removes redundant data while retaining the most pertinent input features. Subsequently, the optimized feature set is processed by the CNN-LSTM architecture. The CNN and LSTM components capture local patterns and sequential dependencies, respectively. The model's effectiveness was assessed by analyzing PV production data and sensor data collected from a real solar power plant over 34 days. This dataset encompassed critical meteorological and electrical parameters, including solar irradiation, ambient and module temperatures, DC power, AC inverter power, and energy efficiency metrics. The experimental results indicate that the proposed NCA–CNN–LSTM model achieves consistently improved performance compared to traditional ML methods and standalone DL models. Although the numerical improvement over the baseline CNN-LSTM model is small, the proposed model provides more efficient feature representation and improved model stability. These features contribute to increased generalization ability and reduced computational complexity.

The proposed model focuses on short-term forecasting scenarios where high temporal resolution is more critical than long-term seasonal variability. The proposed model provides significant practical and economic advantages for grid-connected PV systems. Improved short-term PV forecasting can optimize dispatch and scheduling strategies, decrease reserve requirements, and reduce balancing and imbalance costs. Furthermore, this methodology enhances interaction within electricity markets, thereby fostering cost-effective grid operations. The proposed hybrid DL model, in its entirety, constitutes both a methodological innovation and a practical application, exhibiting considerable promise as a dependable instrument for real-time energy management and decision-making within PV systems.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, [Yasemin Onal], upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tovar M, Robles M, Rashid F. PV power prediction, using CNN-LSTM hybrid neural network model. Case of study: Temixco-Morelos, México. Energies. 2020;13(24):6512. doi:10.3390/en13246512.
2. Bouzerdoum M, Mellit A, Massi Pavan A. A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. Sol Energy. 2013;98:226–35. doi:10.1016/j.solener.2013.10.002.
3. Zhang W, Li Q, He Q. Application of machine learning methods in photovoltaic output power prediction: a review. J Renew Sustain Energy. 2022;14(2):022701. doi:10.1063/5.0082629.
4. Mayer MJ, Gróf G. Extensive comparison of physical models for photovoltaic power forecasting. Appl Energy. 2021;283(C):116239. doi:10.1016/j.apenergy.2020.116239.
5. Wang F, Zhen Z, Liu C, Mi Z, Hodge BM, Shafie-khah M, et al. Image phase shift invariance based cloud motion displacement vector calculation method for ultra-short-term solar PV power forecasting. Energy Convers Manag. 2018;157(6):123–35. doi:10.1016/j.enconman.2017.11.080.
6. Ahmed R, Sreeram V, Mishra Y, Arif MD. A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization. Renew Sustain Energy Rev. 2020;124(4):109792. doi:10.1016/j.rser.2020.109792.
7. Theocharides S, Makrides G, Livera A, Theristis M, Kaimakis P, Georghiou GE. Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. Appl Energy. 2020;268:115023. doi:10.1016/j.apenergy.2020.115023.
8. Sharadga H, Hajimirza S, Balog RS. Time series forecasting of solar power generation for large-scale photovoltaic plants. Renew Energy. 2020;150:797–807. doi:10.1016/j.renene.2019.12.131.
9. Jeong HY, Hong SH, Jeon JS, Lim SC, Kim JC, Park HW, et al. A research of prediction of photovoltaic power using SARIMA model. J Korea Multimed Soc. 2022;25(1):82–91.
10. Samanta S. Forecasting solar power generation in the future by ARIMA approach and stationary transformation. In: Data driven model. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2026. p. 203–20. doi:10.1002/9781394287925.ch9.
11. López Gómez J, Ogando Martínez A, Troncoso Pastoriza F, Febrero Garrido L, Granada Álvarez E, Orosa García JA. Photovoltaic power prediction using artificial neural networks and numerical weather data. Sustainability. 2020;12(24):10295. doi:10.3390/su122410295.
12. Ahmad A, Jin Y, Zhu C, Javed I, Waqar Akram M, Ali Buttar N. Support vector machine based prediction of photovoltaic module and power station parameters. Int J Green Energy. 2020;17(3):219–32. doi:10.1080/15435075.2020.1722131.
13. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: beyond efficient transformer for long sequence time-series forecasting. Proc AAAI Conf Artif Intell. 2021;35(12):11106–15. doi:10.1609/aaai.v35i12.17325.
14. Ma X, Zhang H. Time series forecasting method based on multi-scale feature fusion and autoformer. Appl Sci. 2025;15(7):3768. doi:10.3390/app15073768.
15. Zhao J, Chu F, Xie L, Che Y, Wu Y, Burke AF. A survey of transformer networks for time series forecasting. Comput Sci Rev. 2026;60:100883. doi:10.1016/j.cosrev.2025.100883.
16. Song D, Rehman MSU, Deng X, Xiao Z, Noor J, Yang J, et al. Accurate solar power prediction with advanced hybrid deep learning approach. Eng Appl Artif Intell. 2025;148(12):110367. doi:10.1016/j.engappai.2025.110367.
17. Li J, Rao C, Gao M, Xiao X, Goh M. Efficient calculation of distributed photovoltaic power generation power prediction via deep learning. Renew Energy. 2025;246(1):122901. doi:10.1016/j.renene.2025.122901.
18. Li P, Zhou K, Lu X, Yang S. A hybrid deep learning model for short-term PV power forecasting. Appl Energy. 2020;259:114216. doi:10.1016/j.apenergy.2019.114216.
19. Huang CJ, Kuo PH. Multiple-input deep convolutional neural network model for short-term photovoltaic power forecasting. IEEE Access. 2019;7:74822–34. doi:10.1109/ACCESS.2019.2921238.

20. Wang F, Xuan Z, Zhen Z, Li K, Wang T, Shi M. A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework. Energy Convers Manag. 2020;212(2):112766. doi:10.1016/j.enconman.2020.112766.
21. Shi J, Wang H, Lu G, Wang M, Su K. Secure power forecasting and grid resilience enhancement using a hybrid SVM-RNN model. J Circuits Syst Comput. 2026;35(13):2650092. doi:10.1142/s0218126626500921.
22. Tang Y, Yang K, Zhang S, Zhang Z. Photovoltaic power forecasting: a hybrid deep learning model incorporating transfer learning strategy. Renew Sustain Energy Rev. 2022;162(11):112473. doi:10.1016/j.rser.2022.112473.
23. Chahboun S, Maaroufi M. Principal component analysis and machine learning approaches for photovoltaic power prediction: a comparative study. Appl Sci. 2021;11(17):7943. doi:10.3390/app11177943.
24. Kazem HA, Yousif JH, Chaichan MT, Al-Waeli AHA, Sopian K. Long-term power forecasting using FRNN and PCA models for calculating output parameters in solar photovoltaic generation. Heliyon. 2022;8(1):e08803. doi:10.1016/j.heliyon.2022.e08803.
25. Dimitriadis CN, Passalis N, Georgiadis MC. A deep learning framework for photovoltaic power forecasting in multiple interconnected countries. Sustain Energy Technol Assess. 2025;77:104330. doi:10.1016/j.seta.2025.104330.
26. Kaggle. Solar power generation data [Internet]. 2020 [cited 2026 Jan 1]. Available from: https://www.kaggle.com/datasets/anikannal/solar-power-generation-data.
27. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf Sci. 2012;191(3):192–213. doi:10.1016/j.ins.2011.12.028.
28. Prema V, Bhaskar MS, Almakhles D, Gowtham N, Rao KU. Critical review of data, models and performance metrics for wind and solar power forecast. IEEE Access. 2022;10(1):667–88. doi:10.1109/ACCESS.2021.3137419.
29. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci. 2021;7(3):e623. doi:10.7717/peerj-cs.623.
30. Goldberger J, Hinton GE, Roweis S, Salakhutdinov RR. Neighborhood component analysis [Internet]. 2004 [cited 2026 Jan 1]. Available from: https://papers.nips.cc/paper_files/paper/2004/file/42fe880812925e520249e808937738d2-Paper.pdf.
31. Yang W, Wang K, Zuo W. Neighborhood component feature selection for high-dimensional data. J Comput. 2012;7(1):161–8. doi:10.4304/jcp.7.1.161-168.
32. Wang H, Lei Z, Zhang X, Zhou B, Peng J. A review of deep learning for renewable energy forecasting. Energy Convers Manag. 2019;198(11):111799. doi:10.1016/j.enconman.2019.111799.
33. Murphy K, Schölkopf B, Srivastava N, Hinton G, Krizhevsky A, Sutskever I, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58. doi:10.5555/2627435.2670313.
34. Harris O, Andrews M. Effects and results of dropout layer in reducing overfitting with convolutional neural networks (CNN). World J Adv Eng Technol Sci. 2024;13(2):836–53. doi:10.30574/wjaets.2024.13.2.0584.
35. de Guia JD, Concepcion RS, Calinao HA, Alejandrino J, Dadios EP, Sybingco E. Using stacked long short term memory with principal component analysis for short term prediction of solar irradiance based on weather patterns. In: Proceedings of the 2020 IEEE Region 10 Conference (TENCON); 2020 Nov 16–19; Osaka, Japan. New York, NY, USA: IEEE; 2020. p. 946–51. doi:10.1109/TENCON50793.2020.9293719.
36. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput. 2000;12(10):2451–71. doi:10.1162/089976600300015015.
37. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798–828. doi:10.1109/TPAMI.2013.50.
38. Guyon IM, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82. doi:10.1007/978-3-540-35488-8_1.